Communicating Weight of Forensic Evidence Using a LR: Whose Prior, Whose Likelihoods, and Whom are We Kidding?

Steve Lund & Hari Iyer
Statistical Engineering Division, NIST

Technical Colloquium on Quantifying the Weight of Forensic Evidence, 5/5/2016

Disclaimers

- Viewpoints expressed are my own, are evolving, and do not necessarily reflect the viewpoints of anyone else at NIST (except Hari)

- We don’t claim the viewpoints to be original
  - David Freedman & David Kaye
  - Henry Swafford
  - Cedric Neumann
  - Mark Taper & Subhash Lele

Desirable Properties of Expert Testimony

- What role do experts play in helping jurors make good decisions?
  - convey available information accurately (simplicity helps) and completely (powerfully)

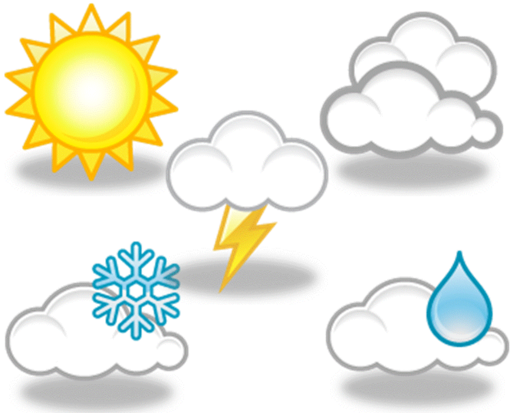
- How to effectively transfer information from one individual to another?

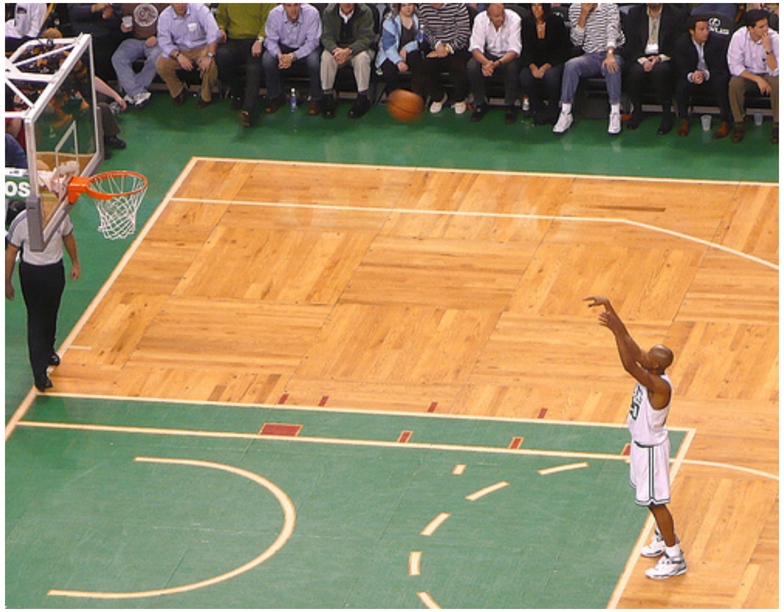
Quantification/Measurement

- Assign an attribute (numerical or non-numerical) to something
- Illuminate relationships (different from, greater than, twice as good….)
  - Attribute in isolation is meaningless
- Relationship clarity follows from **stability** of attribute assignment
  - **Repeatability**: relationships among things measured in a very controlled environment
  - **Reproducibility/Traceability**: relationships among things a broader community has measured (under a broader range of conditions)
- Main point: Communicating a ***single*** number is meaningless without contextual information, possibly a collection of other numbers. In metrology, units provide this context.

Probability

- Declaring a number to have a probabilistic meaning invokes a lot of contextual information, without explicitly providing it





Justifying probability

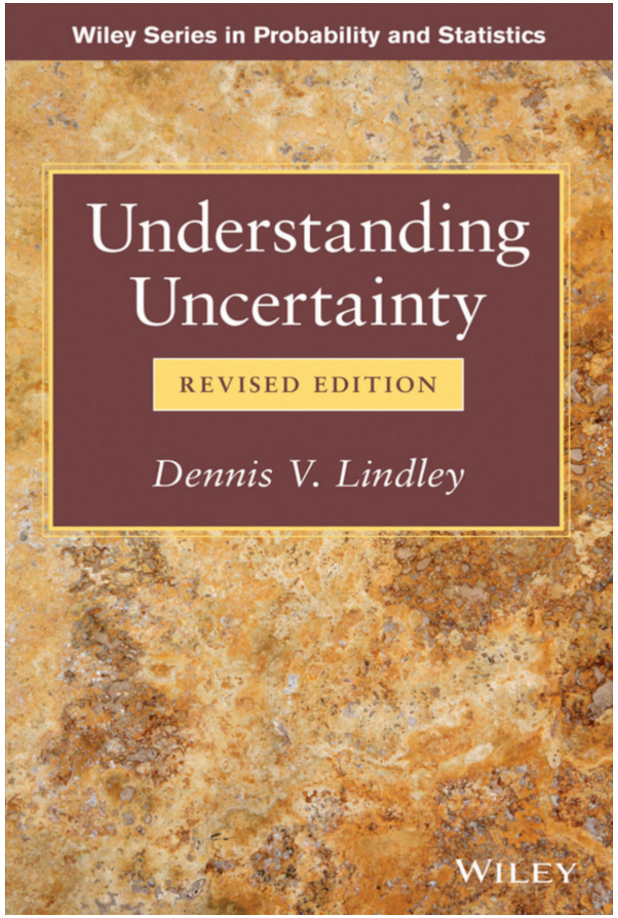
- Chance: Directly understanding the generating mechanism

Not enough information

- Asymptotic Relative Frequency: Having seen a “sufficiently large” collection of its outputs

- Personal belief ?

Subjective Bayes



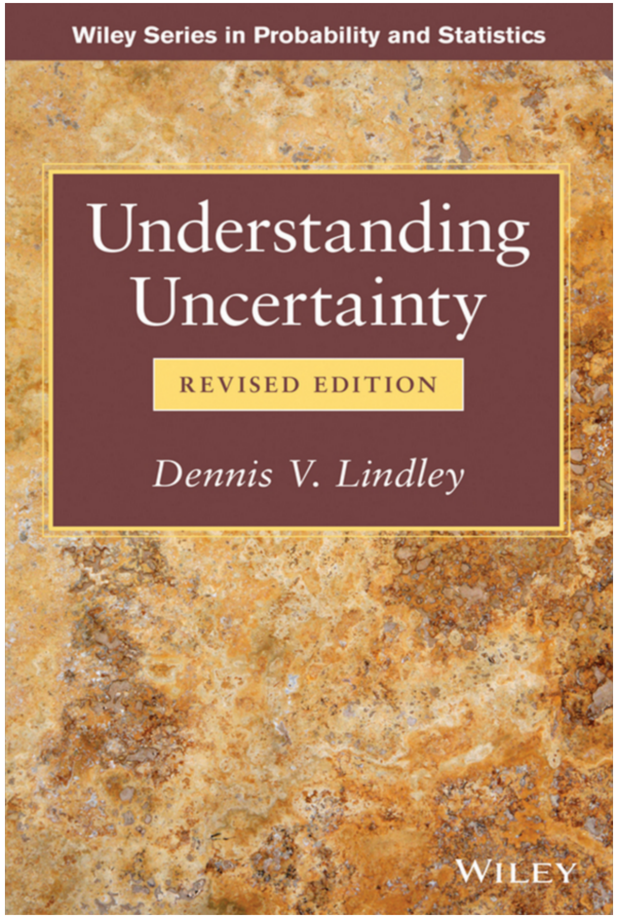
- Beautiful theory developed around Bayes rule:

$$\frac{P[H_p|E]}{P[H_d|E]} = \frac{P[E|H_p]}{P[E|H_d]} \times \frac{P[H_p]}{P[H_d]}$$

Posterior Odds = Likelihood Ratio x *Prior Odds*

- Imprecise probability is an important extension
- Useful guidance for practice
  - We want all TOFs to have a likelihood ratio

Subjective Bayesian Juror



Consider the situation in a court room, where a defendant is charged with some infringement of the law, and suppose it is a trial by jury. There is one uncertain event of importance to the court — Is the defendant guilty of the offence as charged? - which event is denoted by G. Then it is a basic tenet of this book that you, as a member of the jury, have a probability of guilt, $p(G \mid K)$, in the light of your background knowledge K. (There are many trials held without a jury, in which case "you" will be someone else, like a magistrate, but we will continue to speak of "juror" for linguistic convenience.) We saw in §6.6 how evidence E before the court would change your probability to $p(G \mid EK)$ using Bayes rule. The calculation required by the rule needs your likelihood ratio $p(E \mid GK)/p(E \mid G^cK)$, involving your probabilities of the evidence, both under the supposition of guilt and of innocence, G^c. It was emphasized how important it was to consider and to compare evidence in the light both of guilt and of innocence.

Ingredients for LR

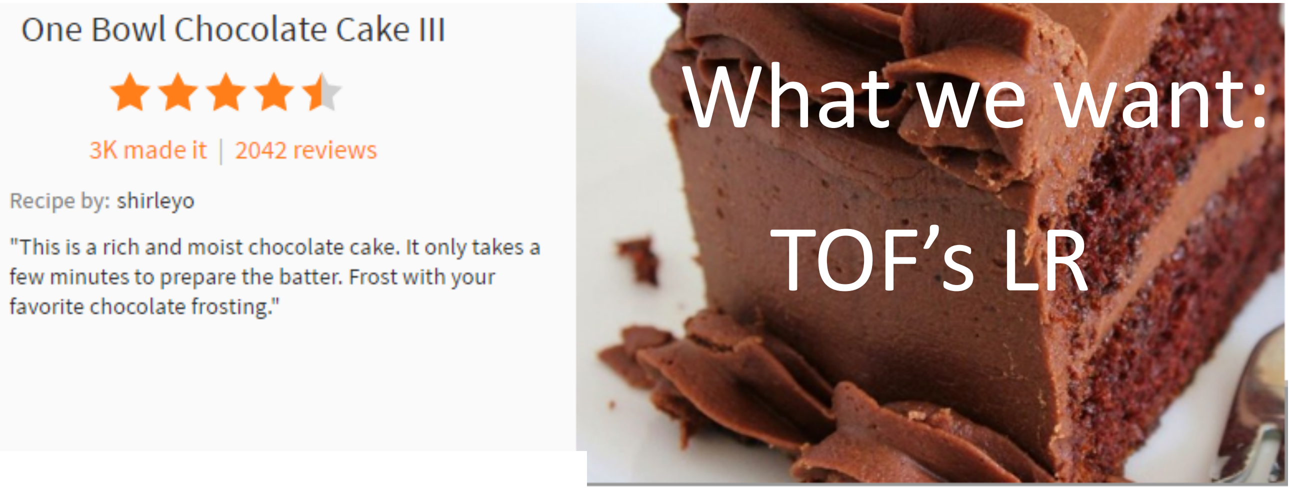
- A set of event sequences (explanations), each providing missing pieces to interpolate between the accepted details of the potential crime and given in sufficient detail to lead to unambiguous conclusions labeled as “guilty” or “not guilty”
- Priors assigned to each explanation (excluding $p(G|K)$, 1 degree of freedom)
  - Relevant “population” (of explanations) is clearly defined as explanations assigned a weight greater than 0
  - ${}^{1}/_{N}$ is not a given
- A likelihood of the evidence under each explanation

THE SETUP

- y = Evidence recovered from the crime scene.
  - $\boldsymbol{S_0}$ is the defendant
  - $\boldsymbol{S_1, \dots, S_N}$ are other potential sources that could have produced y

- x = additional control samples collected from a subset of sources $\boldsymbol{S_0, S_1, \dots, S_N}$

$$LR = \frac{\Pr(y|x, S_0)}{\sum_{j=1}^{N} w_j \Pr(y|x, S_j)}$$


One Bowl Chocolate Cake III
3K made it | 2042 reviews
Recipe by: shirleyo
"This is a rich and moist chocolate cake. It only takes a few minutes to prepare the batter. Frost with your favorite chocolate frosting."
What we want:
TOF's LR

What we have:


Data


Assumptions


1 h
24 serv
On Sale
What's on sale near you.
Target
25 Gran
TARGET
Shopping List
Direc
round pans.

Refractive Index (RI) Example

- 10 RI measurements from crime scene (CS) window and 5 from person of interest (POI):

RI	1.51840	1.51844	1.51846	1.51848	1.51850
# CS	0	2	3	4	1
# POI	1	2	0	1	1

- 2200+ sample mean RIs from different windows
- 49 RIs taken from a single window

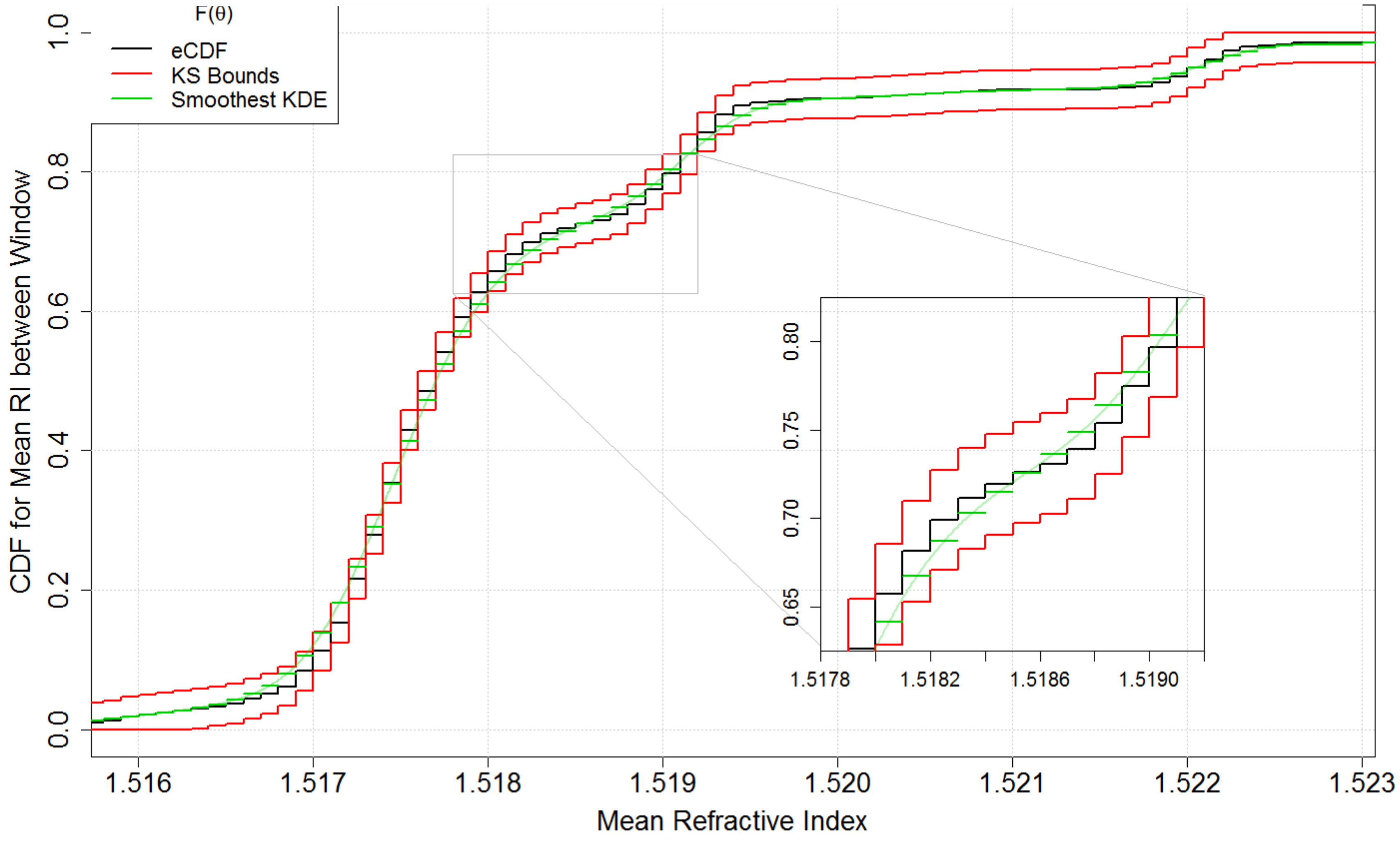
Suppose…

- The 10 RIs from the crime scene window form an i.i.d. sample from population of all RIs in the window
- The 5 fragments from the POI originated from a single window, and RIs form an i.i.d. sample
- The 49 RIs from a single window form an i.i.d. sample
- RI distributions form a location family across windows
- The 2200+ sample means form an i.i.d. sample of location parameters from the relevant population

Suppose…

- The 10 RIs from the crime scene window form an i.i.d sample from population of all RIs in the window
- The 5 fragments from the POI originated from a single window, and RIs form an i.i.d. sample
- The 49 RIs from a single window form an i.i.d. sample
- RI distributions form a location family across windows
- The prior weighted distribution of location parameters in the relevant population is given by the Gaussian kernel density estimate (bandwidth = 0.0001) from the 2200+ sample means [see Aitken and Taroni, 2004]

Empirical CDF for Sample Mean RI
(from 2269 glass fragments from different windows)



H_p: Glass fragments on POI come from CS window, whose average RI was sampled from F

H_d: Glass fragments on POI came from a window other than the CS window, and the average RI for both windows was sampled from F

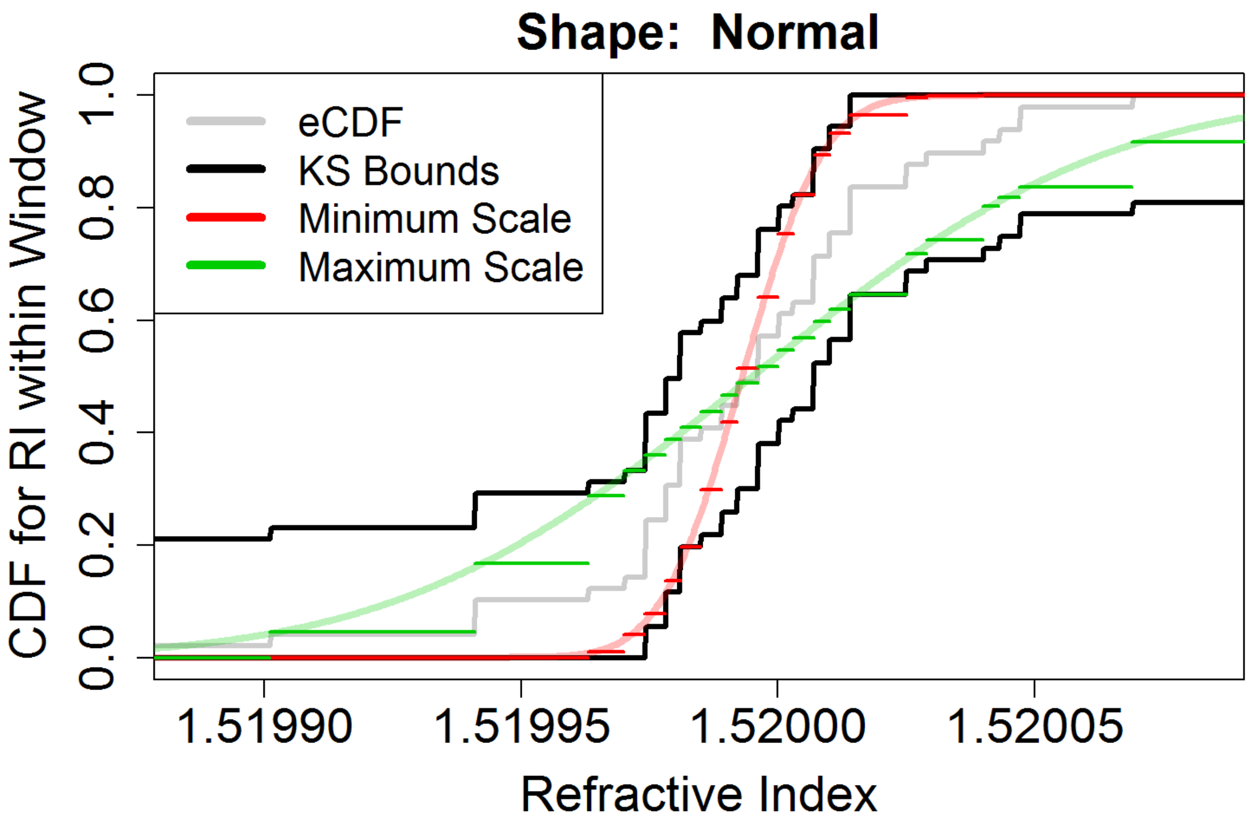
Assume within window RI measurements form location family

Between window location parameter distribution

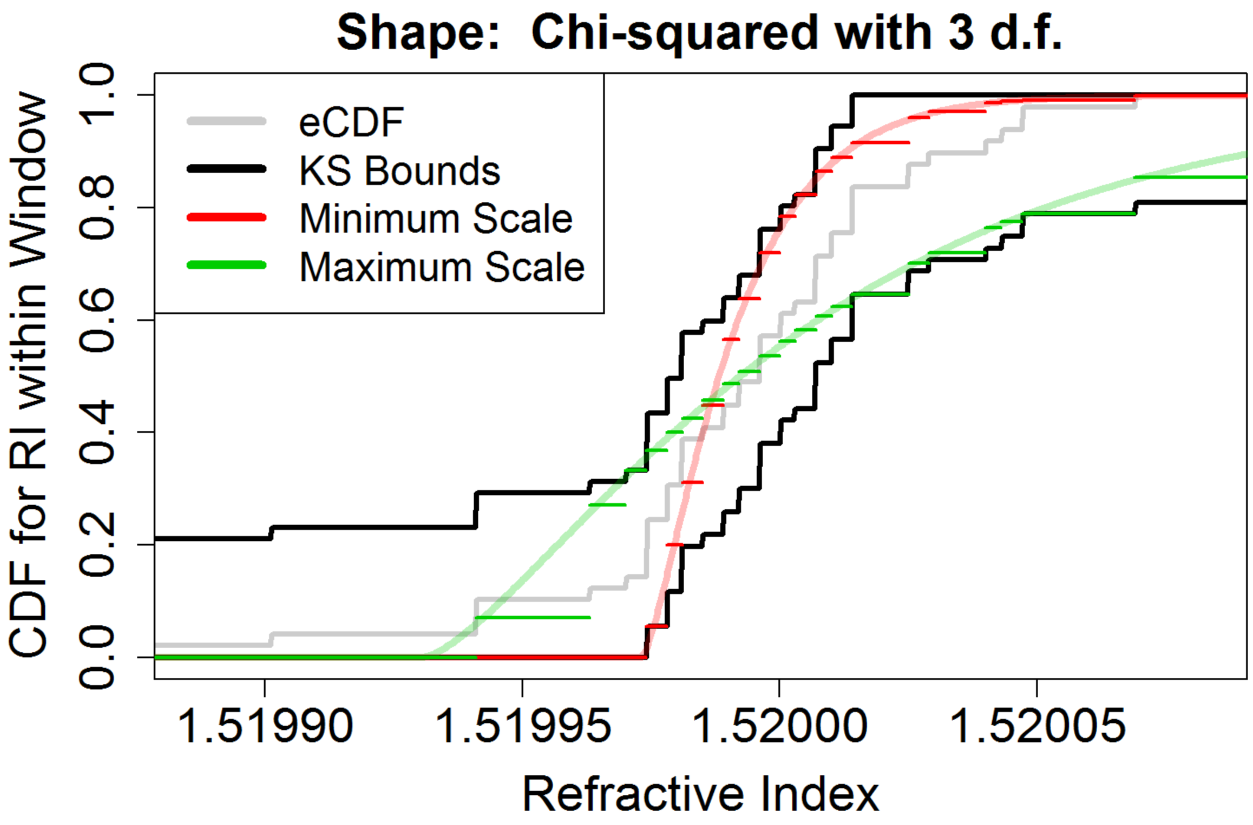
Model: $r_i|\theta \sim G_0(r_i - \theta)$, $\theta \sim F$

$$LR = \frac{\int \prod_{i=1}^{15} g_0(r_i - \theta)dF(\theta)}{\int \prod_{i=1}^{5} g_0(r_i - \theta)dF(\theta) \int \prod_{i=6}^{15} g_0(r_i - \theta)dF(\theta)}$$

Illustrative exercise: Examine range of LR over choices of g_0 that fall within 95% KS Confidence band of available data



LR Range: 65 to 196



LR Range: $\frac{3}{1{,}000{,}000}$ to 22

What if RI distributions vary in ways other than location across windows??

What if the sampled windows aren't representative of my relevant glass population?

What if there's an unknown correlation structure among the RI??

What if glass recovered from POI came from more than one place??

Justifying probability

- Chance: Directly understanding the generating mechanism

Not enough information

Not enough informati

- Asymptotic Relative Frequency: Having seen a "sufficiently large" collection of its outputs

"We do not care what you believe, we barely care what we believe, what we are interested in is what you can show."

- Personal belief ?

- Taper and Lele, *Evidence, Evidence Functions, and Error Probabilities*, from the book *Philosophy of Statistics*. (2011)

What comes out of an LR computation?

- ***Maybe*** the subjective LR from Bayes' formula for the analyst

- What does the TOF ***do*** with it?
  - It's a mysterious number produced by an algorithm (with fuzzy inputs) and reported to have good discriminating efficiency
  - … i.e. a ***score***, the appropriate interpretation of which is neither self-evident nor uniquely known (contextual information is required)

What would you rather testimony do?

- Focus on ***actual*** information, centered around the ***case***
- Explain what was done and why
- Describe data that illustrates both an event and its meaning
  - Clearly describe how data was obtained and be open about its limitations
- Avoid claims (probabilistic or otherwise) that aren’t supported by demonstrable data
  - Models are imaginary, but influential

Selected References

1. Ramsey, F. P. (1931). Foundations of Mathematics and Other Logical Essays (Early definition of personal probability using betting ideas)
2. de Finetti, B. (1974), Theory of Probability, Vol. 1. New York: John Wiley and Sons.
3. David Freedman (1995). Some Issues in the Foundation of Statistics, Foundations of Science.
4. Savage, L.J. (1954), The Foundations of Statistics. New York: John Wiley and Sons (second edition, 1972, New York: Dover).
5. Kadane, J. B. (2011). Principles of Uncertainty, CRC Press.
6. Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm, Chapman & Hall.
7. Aitken, C. G. G. and Taroni, F. (2004). Statistics and the Evaluation of Evidence for Forensic Scientists, Second Edition, John Wiley and Sons.
8. Aldrich, John. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922, Statistical Science 1997, Vol. 12, No. 3, 162-176. (Notion of Likelihood & Max. Likelihood)
9. Lindley, D. V. (1977). A Problem in Forensic Science, *Biometrika,* Vol. 64, No. 2, pp. 207-213
10. Tribe, Laurence. H. (1971). Trial by Mathematics: Precision and Ritual in the Legal Process, Harvard Law Review.
11. Finkelstein, M. O. and Fairley, W. B. (1970). A Bayesian Approach to Identification Evidence, Harvard Law Review.
12. Peter Walley, (1991). Statistical Reasoning with Imprecise Probabilities, Springer.

Thank you!

What was done and why?

- How were similarity and quality metrics chosen?
  - What data was used and where is it from?
  - ROC curves

Describe data to illustrate an event

- What is the quality of the questioned? Of the known(s)?

- What is the similarity between the questioned and known(s) attributed to the POI?
  - Subjective impression
  - Algorithm score / LR
  - Classifier / Categorical conclusion from expert / verbal scale

… and its meaning

Part 1: Specific Source

- What is the similarity between the questioned and other knowns, not attributed to the POI?
  - Less useful when POI has been chosen as result of database search
  - Otherwise, may indicate some level of rarity to degree of correspondence; *i.e.* the POI is the best match in a crowd

… and its meaning

Part 2: Common Source

- In controlled cases, what similarities have occurred between a questioned and mated known(s) under “comparable conditions?”

- … and for non-mated knowns?
  - When collection of sources has hierarchical/clustered structure, breakdown the within/between.   E.g.
    - consecutively manufactured
    - same size, make, model
    - different make, model

“Comparable Conditions”

- Stricter definitions mean fewer observations (less information) or more $$
- How much pooling should we do?
    - How consistent are similarity distributions across sources that span time, location, race, brand, etc.?
    - How consistent are similarity distribution across various combinations of questioned and known qualities? Time interval/exposure between collection of questioned and known?
    - ROC curves

Final Remarks

- To be valuable, information requires context
  - Potential vs. realized value of evidence
- We may not be able to afford fully realizing the potential value of evidence
  - Desire to imagine we had all the data (distributions as specified model)
  - Extrapolate far past what can be empirically shown (opening Pandora's box)
- Caution!
  - Interpretations are sensitive to distribution tails
  - Ground truth is generally unknown, removing guardrails, and false confidence can destroy lives
- To improve real value of evidence, set up data bases, develop quality and similarity metrics, and focus on effective descriptions

Biometrics by Algorithm

1. Acquire
2. Process
3. Act

Biometrics by Human

1. Acquire
2. Process
3. Interpret
4. Act

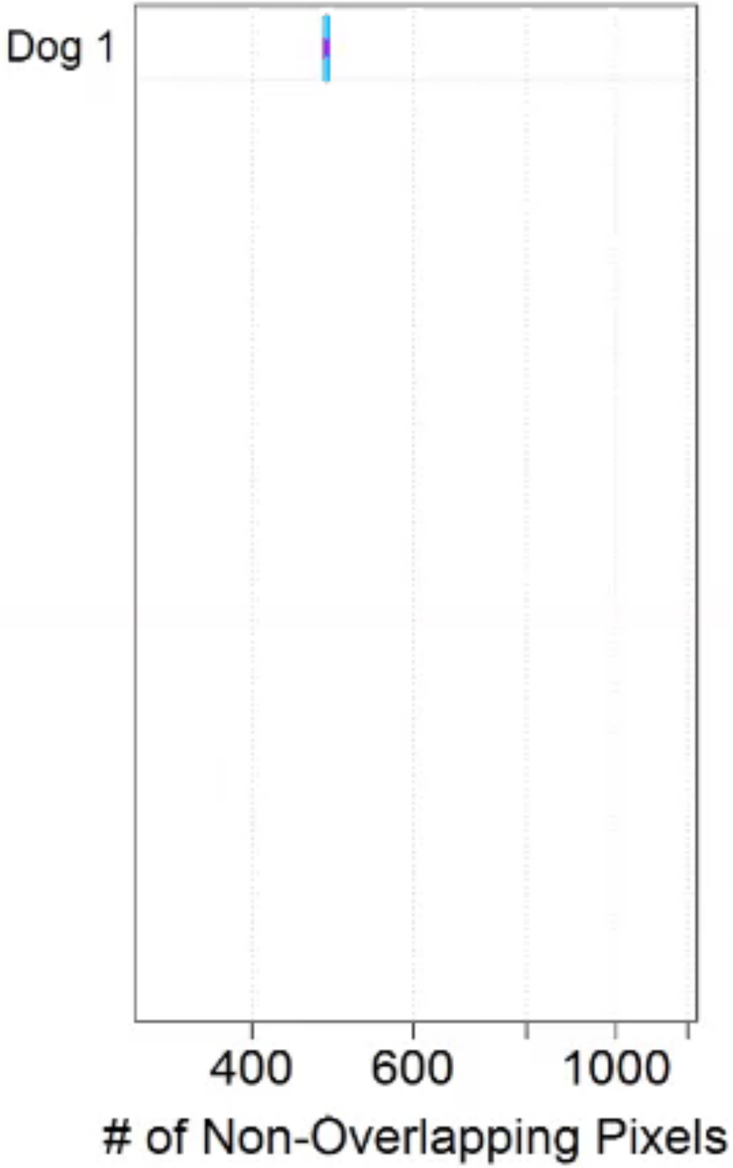
Expert & Trier of Fact

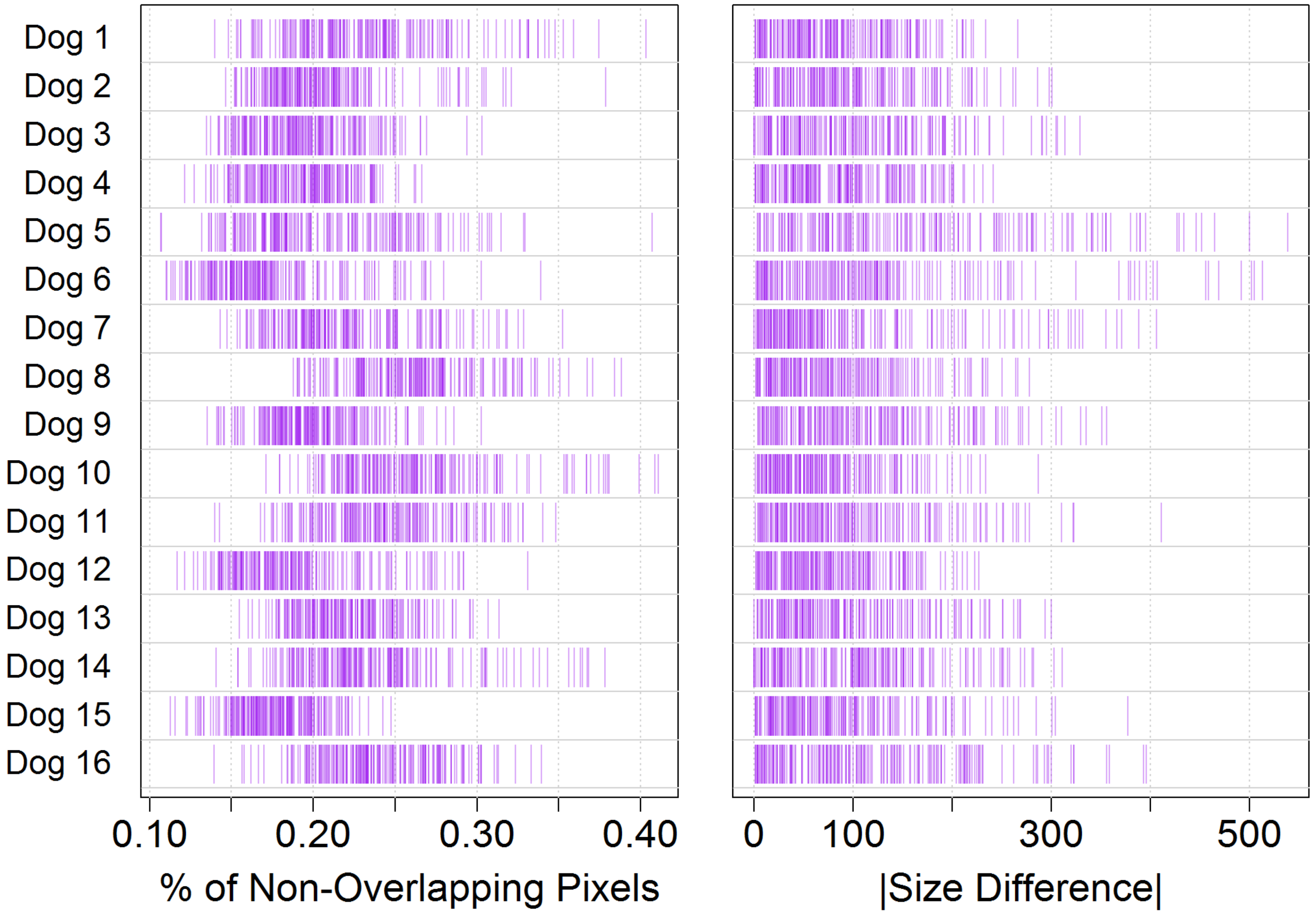
1. Acquire
2. Process
3. Interpret
4. Act

Transfer: Report / Receive

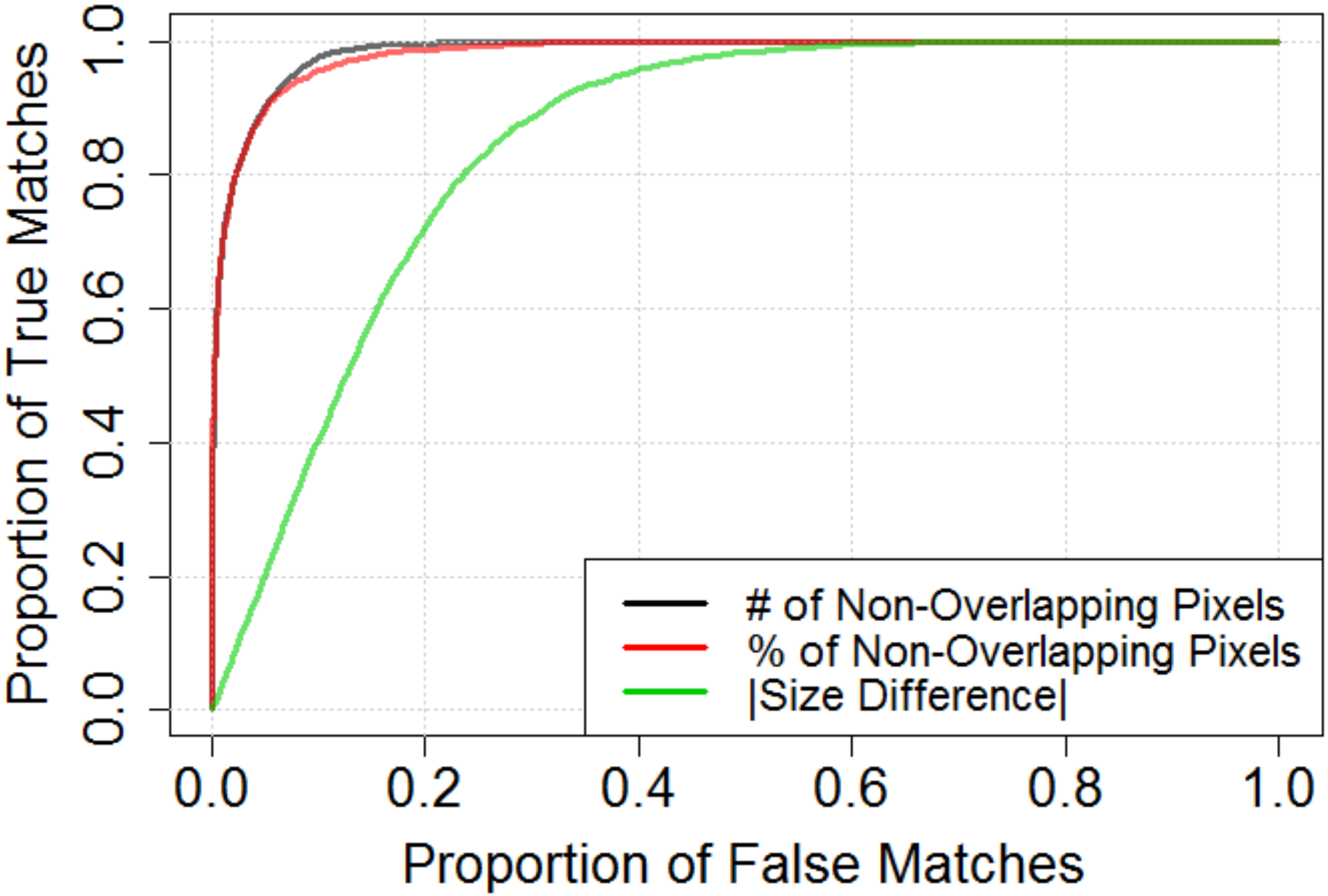
Transfer: Report / Receive


743
Dog 1, Print 1
Dog 1, Print 2

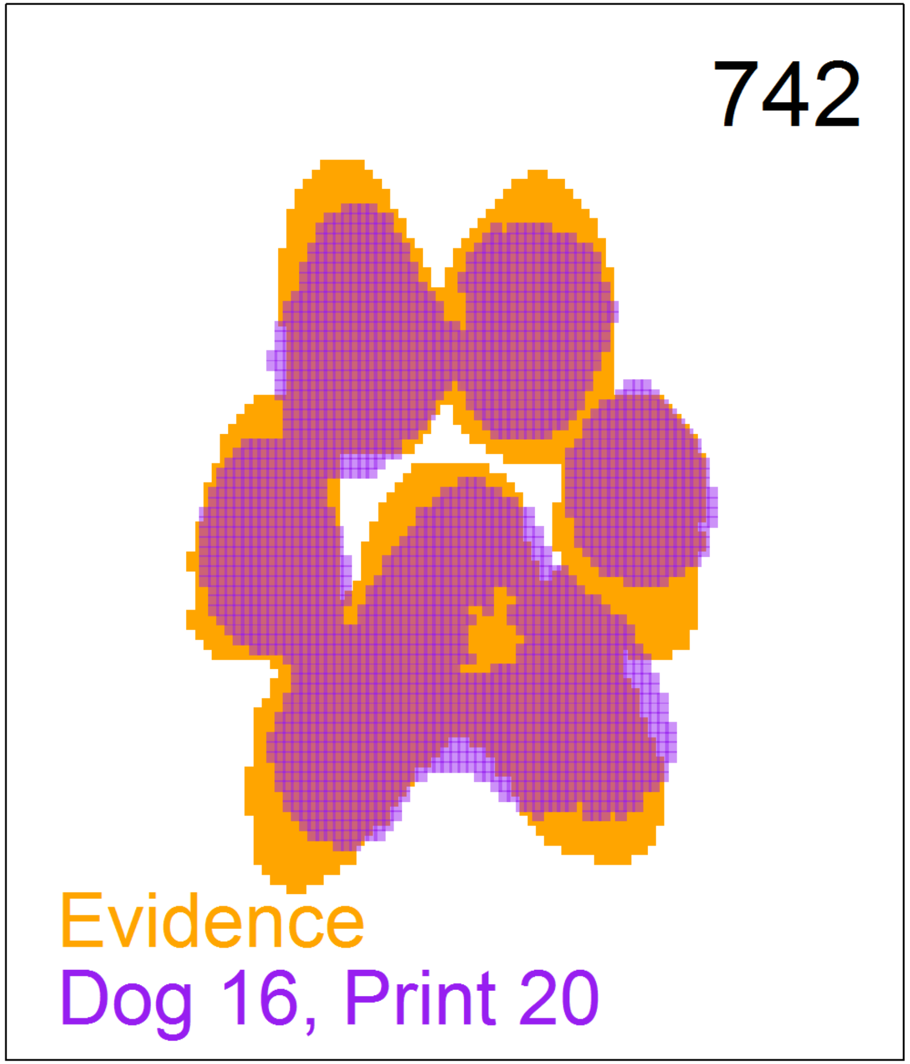

Dog 1
400
600
1000
# of Non-Overlapping Pixels



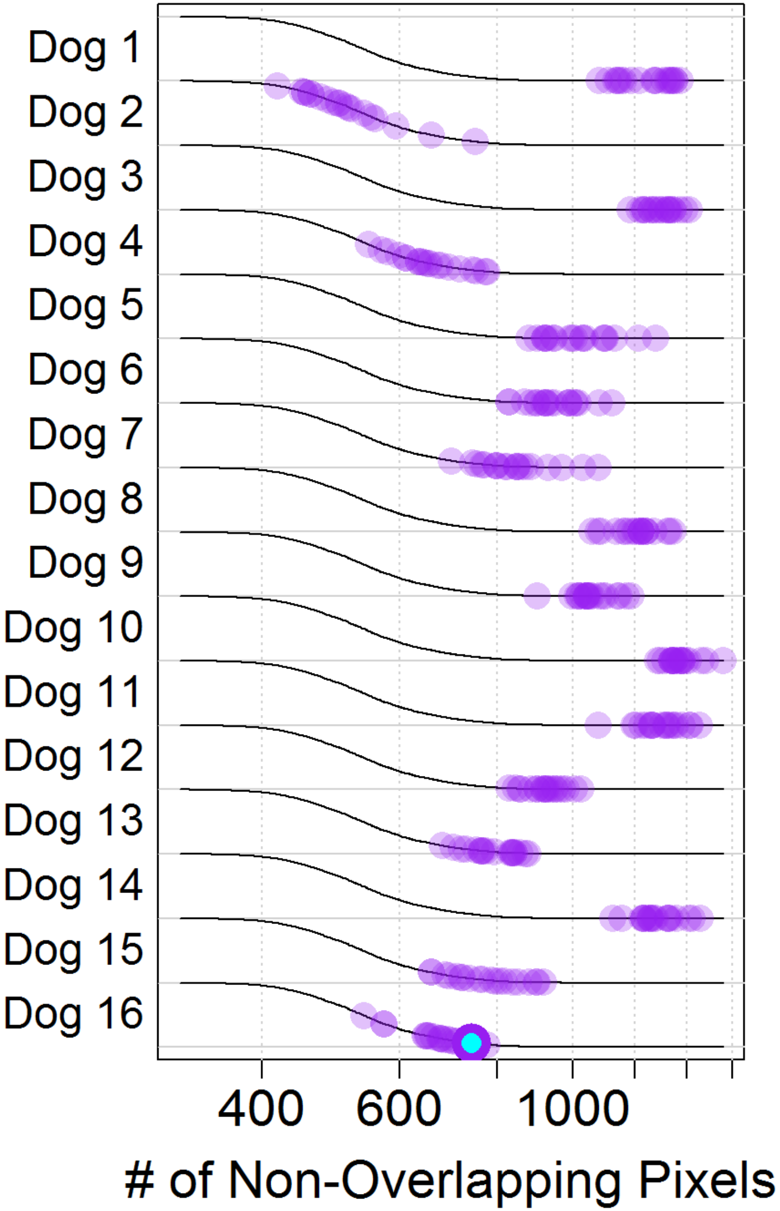
Dog 1
Dog 2
Dog 3
Dog 4
Dog 5
Dog 6
Dog 7
Dog 8
Dog 9
Dog 10
Dog 11
Dog 12
Dog 13
Dog 14
Dog 15
Dog 16
0.10
0.20
0.30
0.40
% of Non-Overlapping Pixels
0
100
300
500
|Size Difference|



Proportion of True Matches
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of False Matches
# of Non-Overlapping Pixels
% of Non-Overlapping Pixels
|Size Difference|

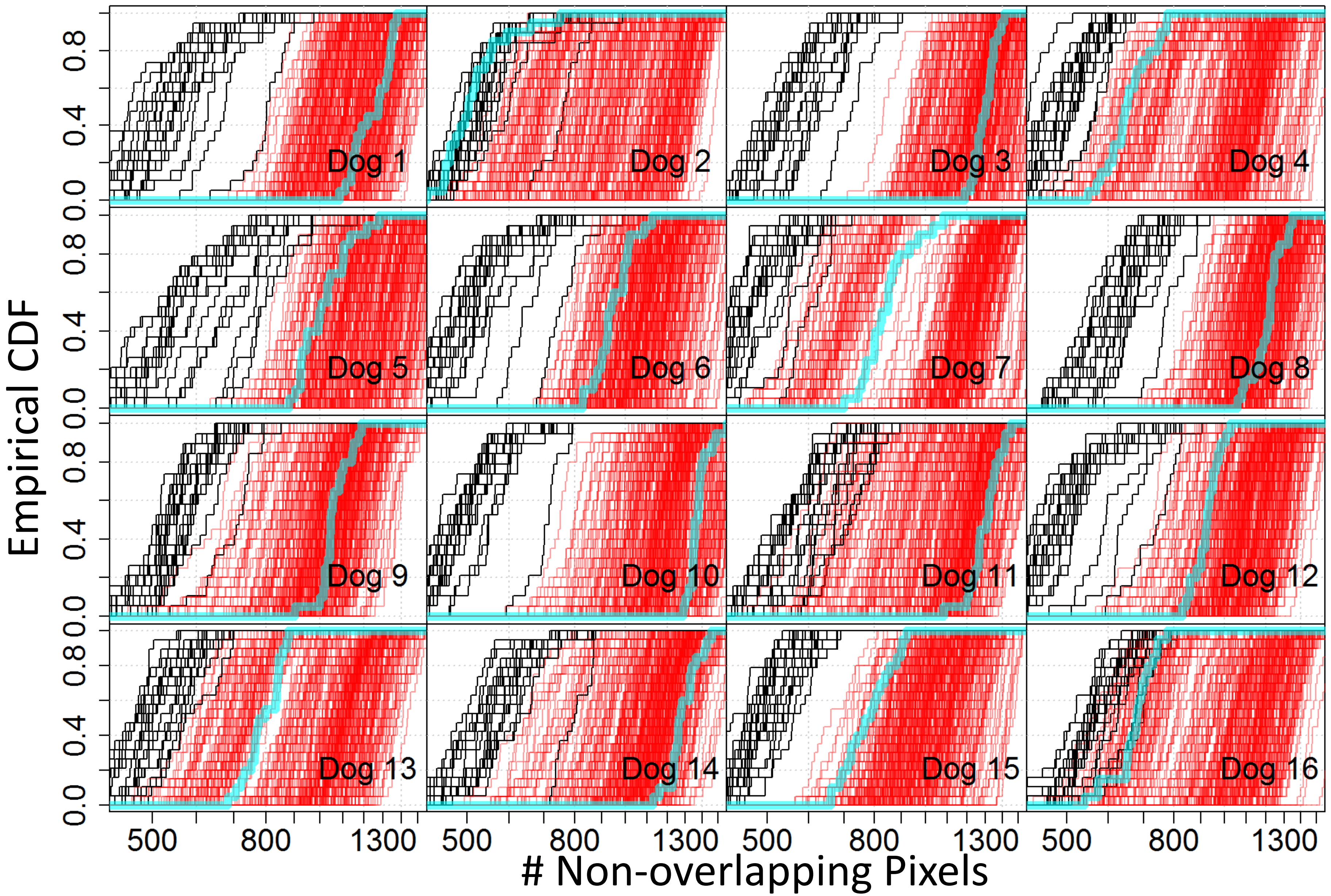


742
Evidence
Dog 16, Print 20

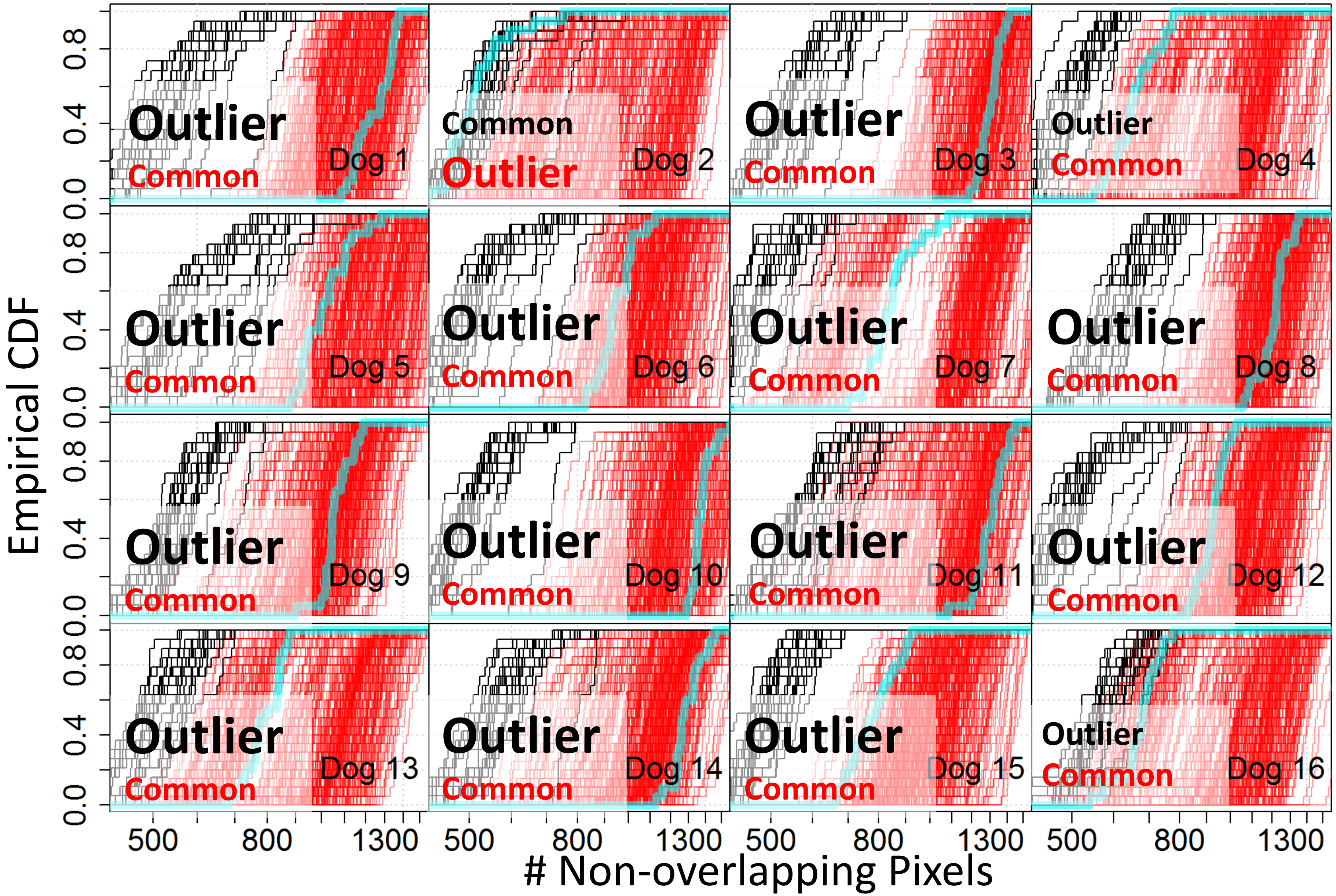


Dog 1
Dog 2
Dog 3
Dog 4
Dog 5
Dog 6
Dog 7
Dog 8
Dog 9
Dog 10
Dog 11
Dog 12
Dog 13
Dog 14
Dog 15
Dog 16
400
600
1000
# of Non-Overlapping Pixels

eCDF from comparisons with the evidence vs eCDFs from known comparisons involving own paw print set



eCDF from comparisons with the evidence vs eCDFs from known comparisons involving own paw print set



eCDF from comparisons with the evidence vs eCDFs from all known comparisons